COVER PAGE
INVITED PAPER MANUSCRIPT
15th INTERNATIONAL CONFERENCE of the AUDIO ENGINEERING SOCIETY

Headphone and head-mounted visual displays for virtual environments

Durand R. Begault
San Jose State University
Stephen R. Ellis
Elizabeth M. Wenzel
Human Factors Research & Technology Division
NASA Ames Research Center
Mail Stop 262-2
Moffett Field, California 94035-1000
Email db@eos.arc.nasa.gov

This paper will describe some of the intersensory phenomena that arise during operator interaction within combined visual & auditory virtual environments. Conjectures regarding audio-visual interaction will be proposed.

INTRODUCTION. A realistic auditory environment can contribute to both the overall subjective sense of presence in a virtual display, and to a quantitative metric predicting human performance. Here, the role of audio in a virtual display and the importance of auditory-visual interaction are examined. Conjectures are proposed regarding the effectiveness of audio compared to visual information for creating a sensation of immersion, the frame of reference within a virtual display, and the compensation of visual fidelity by supplying auditory information. Future areas of research are outlined for improving simulations of virtual visual and acoustic spaces.

DEFINITIONS OF VIRTUAL EVIRONMENT DISPLAY FIDELITY. Two commonly used approaches to measuring fidelity in virtual displays involve objective and subjective measures to assess operator performance [1]. Subjective measures tend to be much more convenient but are prone to individual variation. Training and introduction of hierarchical ratings such as the Cooper-Harper scale of aircraft controllability [2] may address this problem. A scale adapted from this rating scale was applied by Ellis, et al. [3] for subjective estimates of the controllability of a virtual cursor used by subjects.

The alternative to subjective measurements of presence is to look specifically at operator performance within a complex task, such as driving a car or operating some other type of machinery. Perceptual simulation criteria can be established via psychophysical investigations, where performance and perception can be measured statistically. It is possible to test, for instance, the effect of diminishing update rates or sonic fidelity and evaluate the change in performance. One example of this is a "tracking task" described in Ellis, et al. [3]; see Figure 1. Another example is those studies that investigate degradation in auditory localization ability as a function of the fidelity of the simulation [4].

AUDITORY PRESENCE VERSUS AUDITORY VIRTUALIZATION. In attempting to describe levels of virtual acoustic simulation, one can define *auditory presence* to mean the ability to subjectively convince the user of their presence in an auditory environment. On the other hand, *auditory virtualization* refers to the ability to simulate an acoustic environment such that performance by the listener is indistinguishable from their performance in the real world.

It is usually necessary to incorporate a high level of fidelity in a simulation to achieve auditory virtualization. Note however that performance is not necessarily tied to "realism." Equating "auditory presence" with "realism" in audio is primarily a falsehood that originated in the audio industry's marketing of sound reproduction equipment; like television, any loudspeaker listening test is in fact a preference choice between different versions of reality, not a measure of reality itself. Begault [5] describes the possibilities of simulating a real (remembered) environment (e.g., a violin concert at Carnegie Hall); unreal possibilities (the violinist flies around the head like an angry insect) or what might be called "auditory morphing" (the violinist's environment changes from Carnegie Hall to a polar ice field). How does one assess the most realistic "morphing"? Note also that if one uses a truly veridical simulation of the actual sound pressure level of gunfire or explosives in a game or training scenario with a military theme, the user may in fact become temporarily or even permanently deafened to any other auditory stimuli.

One aspect of presence that has been described is "immersion" [6]. Given the definition of immersion as being surrounded, it is easier to convey this sensation via audio than with vision. Both natural and loudspeaker-generated sound can create an immediate sense of being surrounded without exploratory head movement, unlike visual stimuli within a helmet-mounted display. If immersion means a sense of being "awash in stimuli," sound has a natural advantage from a physical standpoint because visual wavelengths are small relative to sound wavelengths. Low frequencies immerse a listener; for instance, a 50 Hz wavelength is around 6.5 meters long, much wider than a head or the width of the field of vision within a visual display. The omni-directional nature of

acoustics may favor audio as a tool for creating immersive displays, compared to the restrictions of the visual field that are only overcome by head motion.

LEVELS OF FIDELITY. At a minimum, virtual reality systems usually involve visual interaction on a variable level of complexity. Ellis [1] has schematized increasing levels of sensory information and environmental control as shown in the left side of Figure 2. Note that the level of visual information contained within the display drives the hierarchy; audio only appears at the "virtual environment" level. This is both recognition of the dominance of vision over other perceptual senses ("visual dominance") as well as perhaps a reflection on virtual reality's tie to the evolution of flight simulation. The ability to fly an aircraft, navigate through an unknown space, or manipulate an object is driven by visual considerations.

A similar hierarchy of levels can be applied to spatial auditory stimuli, as seen at the right side of Figure 2. At the lowest level of fidelity, basic localizing functions in the environment are enabled via lateralization, where sound sources are positioned to the left or right based on interaural intensity and time-of-arrival differences, combined with a simple distance perception based on sound pressure level. Psychophysical studies support the fact that localization is most accurate for azimuth, and that the primary cue to distance is sound pressure level. This is the basis of "stereo" systems and the ability to place virtual sound images between two speakers. As with a pictorial representation of space, where scale is arbitrary, the relative placement of virtual acoustic images can be successfully communicated via simple communication systems that need simulate only rudimentary spatial cues. Immersion is unnecessary in order to hear a spatial distribution of sound from, e.g., a popular recording.

Hearing in 3-D dimensions (azimuth, elevation and distance) with proprioceptive feedback from a head tracking device constitutes a secondary level of fidelity. These cues are primarily derived from measurements of head-related transfer functions (HRTFs- see [5] for further information). The lack of environmental information from diffuse field reverberation can cause perceptual errors such as failures of externalization. Early 3-D sound systems were essentially renderings of the sound field within an anechoic chamber, with only overall sound pressure used as a cue to distance.

The addition of cues to an environmental context ("auralization") of the complete acoustic environment involves the interaction of the sound source with reflective and absorptive materials. At this level of fidelity, cues to distance can be derived from reverberation, as can cognitive cues to the nature of the specific space. Consequent cues to the width and "spaciousness" (sense of being surrounded or immersed) are also obtained from reverberation modeling in virtual environments. It is also feasible to supply cues to sounds in the range of approximately 20-100 Hz that are both felt as vibration and heard as sound, perhaps providing a more "visceral" sense of immersion.

In all but the most minimal of simulations, the addition of interactivity can dramatically improve the fidelity of a virtual environment. For example, enabling head motion greatly decreases front-back confusion errors in anechoic simulations compared to static (fixed head and fixed sound source) situations [7]. There is even some evidence that simple interaural time differences tied to head motion can reduce front-back confusion errors to some degree [8]. Just as moving the head causes dynamic changes for a fixed source, a moving source will cause dynamic changes for a fixed head. However, Wightman and colleagues have shown that source motion per se does not appear to reduce localization errors unless the motion is under the control of the listener, e.g., either via head motion or keyboard control.

AUDITORY-VISUAL INTERACTION IN VIRTUAL ENVIRONMENTS. One of the principal advantages of sound in a virtual environment is situational awareness; cueing or communication streams can be provided from all around a listener, while visual communication is confined to the specific viewing angle and line-of-sight. A widely discussed, although seldom proven, advantage of sound is the notion that the perceived quality of a visual display will improve with suitable sound cues. Brenda Laurel described that, when producing video games, she noticed that "really high-quality audio will actually make people tell you that games have better pictures, but really good pictures will not make audio sound better; in fact, they make audio sound worse [9]." Laurel's anecdotal experience has recently been corroborated experimentally. Russell Storms, a Ph.D. candidate at the Naval Post Graduate School in Monterey, California, recently completed a study on the interaction of three different resolution levels of audio and visual displays [10]. A statistically significant effect was found showing that judgments of a computer-presented visual image are biased toward a higher level of perceived quality with the addition of either medium or high quality sound.

Feedback is another advantage of audio in a virtual display. For instance, the VPL Data Glove worked by analyzing a combination of detected finger positions and then using the software to match them to sets of predefined gestures (e.g., a peace sign, pointing with the index finger, squeezing hand as a fist). It can be tricky at first to adapt one's hand actions to recognized gestures because individual variation in hand size and movement must match a generalized gesture. But if auditory feedback is supplied upon successful recognition, it aids the user in knowing when the action has been completed successfully.

NASA's VIEW system of the 1980s is an early example of how 3-D audio can serve as a source of tactile and performance feedback. Wenzel, *et al.* [11] created an

audio system that provided symbolic acoustic signals that matched the semantic content of operator actions. The sounds were not only spatialized in 3-D dimensions for conveying situational awareness information but were also interactively manipulated in terms of pitch, loudness and timbre. Audio feedback that supplemented or replaced force feedback could also be represented as a continuum by changing one or more sound parameters in correspondence to the force's intensity.

CONJECTURES AND FUTURE WORK. Systematic investigation of the performance improvement in virtual displays with corresponding audio cueing is a topic of future research by the present authors. Though the application areas for the use of spatialized sound are generally dynamic, very limited information on the perception of these dynamic properties is available. Much of the literature on the importance of visual cues in auditory localization is concerned with static cues, in particular, the so-called "ventriloquism" effect, whereby the apparent position of an auditory object can be dominated by the presence of a correlated visual object [12]. Since dynamic cues present a greater challenge to the perceptual system than static ones, we need to ascertain how the visual and auditory cues interact, in order to calibrate the cues and determine their relative utility.

There is also little experimental evidence to date that poor display fidelity can be compensated by good interactivity, both within and across modalities. Given the shift in subjective evaluations of visual imagery with the presence of audio, it may be possible to relax computational requirements for visual rendering. Specifically, it may be possible to compensate for lower update rates or greater latency within a visual display by simultaneously supplying high-fidelity spatial audio.

Combining virtual acoustic and visual cues may also improve subjects' ability to track and localize virtual sound sources as they move through a room or an environment. The frame of reference used when tracking a set of visual-auditory objects in virtual space may also be important. For example, it has been informally observed that an excocentric frame of reference for moving virtual objects, with sources placed only to the front of an observer, are easier to track than an egocentric perspective where the user is surrounded by the sources (see Figure 3).

Finally, there is a symbiosis between the goals of high-fidelity simulation of acoustic spaces and localization accuracy. To achieve sound source externalization (and consequently an optimal level of auditory fidelity), simulation of a diffuse field is necessary [4, 13]. It is possible, but not yet definitively proven, that knowledge of the acoustic features of a space may aid in localization, or even that there is an optimal acoustic environment for supporting localization. Moreover, the change in an auralized pattern of reflections must be "plausible" to prevent a release from echo suppression and a consequent degradation in externalization of sound sources [14]. Accurate rendering of the acoustic features of the HRTFs of the listener and of the reflected environment is expensive, but localization accuracy may be optimized under such conditions. Ultimately, future research may be able to establish engineering guidelines that allow a minimal amount of computation for a particular simulation scenario [15, 16].

REFERENCES

[1] Ellis, S. R. (1995). Presence of mind: a reaction to Sheridan's "Further musings on the psychophysics of telepresence. *Presence*, 5, 247-259

[2] Cooper, G. E. & Harper, R. P. Jr. (1969) The use of pilot ratings in the evaluation of aircraft handling qualities. NASA Technical Report TN-D-5153.

[3] Ellis, S. R., N. S. Dorighi, N. S., Menges,B. M., Adelstein, B. D., and Jacoby, R. H. (1997). In search of equivalence classes in subjective scales of reality *Proceedings of HCI International '97 Elsevier, Amsterdam* pp. 873-876.

[4] Begault, D. R. (1992). Perceptual effects of synthetic reverberation on three–dimensional audio systems. *Journal of the Audio Engineering Society*, 40, 895–904.

[5] Begault, D. R. (1994). *3-D Sound for Virtual Reality and Multimedia.* Cambridge, MA.: Academic Press Professional.

[6] Witmer, B. G. and Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence*, 7, 225-240

[7] Wenzel, E. M. (1995) "The relative contribution of interaural time and magnitude cues to dynamic sound localization." *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,* New Paltz, NY, October 15-18, Piscataway, NJ: IEEE Press..

[8] Wenzel, E. M. (1996) "Effectiveness of Interaural Delays Alone as Cues during Dynamic Sound Localization." [Abstract] *Acoustical Society of America,* 100, 2608.

[9] Tierney, J. (1993, September 16). Jung in Motion, Virtually, and Other Computer Fuzz. *The New York Times*, B1–4.

[10] Storms, R. L. (1998) Personal Communication.

[11] Wenzel, E. M., Fisher, S., Stone, P. K., and Foster, S. H. (1990). A system for three–dimensional acoustic "visualization" in a virtual environment workstation. In *Proceedings of Visualization '90 Conference.* New York: IEEE Press.

[12] Welch, R. B. (1978). *Perceptual modification: Adapting to altered sensory environments.* New York: Academic Press.

[13] Plenge, G. (1974). On the difference between localization and lateralization. *Journal of the Acoustical Society of America*, 56, 944–951

[14] Blauert, J. (1997). *Spatial hearing. The Psychophysics of Human Sound Localization* (Revised edition). Cambridge, MA: MIT Press.

[15] Begault, D. R. (1997). Virtual acoustics: Evaluation of psychoacoustic parameters. *Third International Workshop on Human Interface Technology IWHIT '97*, Aizu, Japan.

[16] Begault, D. R., Wenzel, E. M., Tran, L. L., & Anderson, M. R. (1998). Octave-band thresholds for Modeled Reverberant Fields (preprint 4662). *Audio Engineering Society 104th Convention*, Amsterdam.

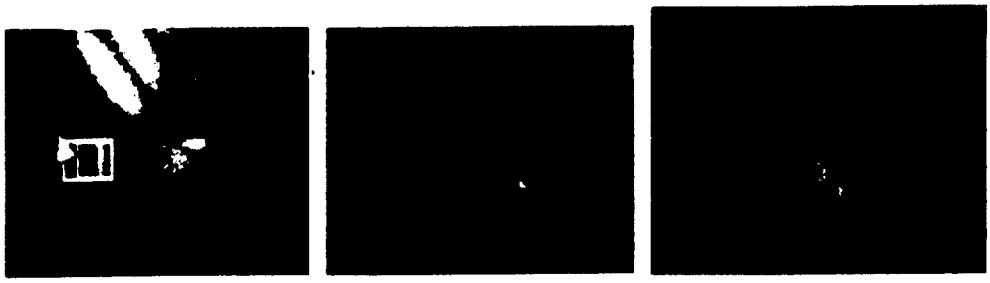

Figure 1. The left panel shows a subject performing the 3-D tracking task by attempting to keep the tetrahedron inside the moving cube. The subject's actual view through the head- mounted display is represented by a screen image in the middle panel. A closer view is on the right. (From [3]).

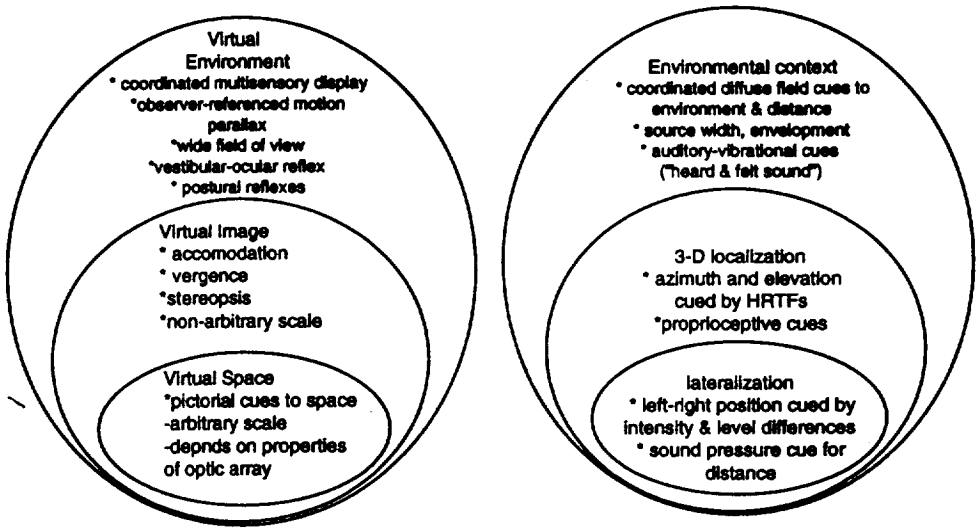

Figure 2. Heriarchies of visual and auditory fidelity levels compared (left diagram from [1]).

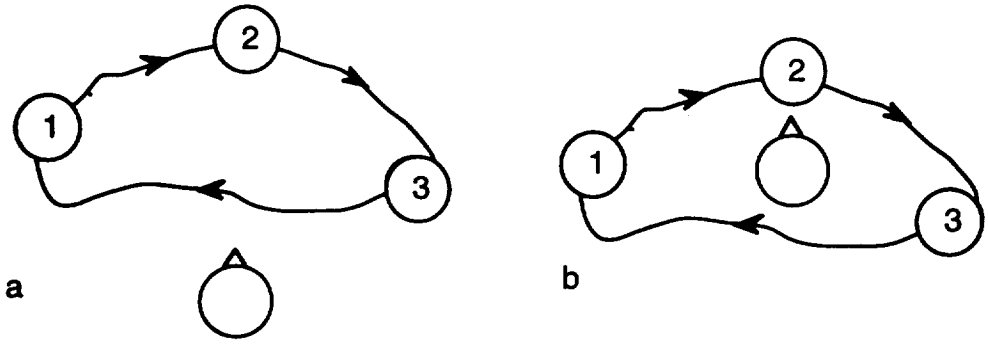

Figure 3. Exocentric (left) vs egocentric (right) comparison of moving sound sources relative to a listener. Numbers refer to audio-visual virtual "objects."